Measure of performance for kernel clustering

Farag Shuweihdi* & Charles C. Taylor

Department of Statistics, University of Leeds

1 Introduction

Kernel density estimation (Wand, 1995), considered as a tool of non-parametric density estimation, has been used for exploring featurs of data set such as the location of modes. Because of this, it is interesting to employ kernel density estimation methods to detect clusters(classes). This approach is based on locating the modes (local minima), the true density, f, has continuous second dervatives. The kernel estimator, $\hat{f}_h(x)$ is given by

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{1.1}$$

where, $X_1, \cdots, X_n$ is a random sample taken from a continuous univariate density, f, $K(x)$ is called kernel and usually assumed to be a probability continuous density function and symmetric about zero with variance, σ_K^2, such as unimodal normal density. The smoothing parameter, h, in equation (1), controls the amount of smoothing of the kernel estimator. Each mode of the kernel density estimator represents a region of high density in the data set. Thus clusters can be discovered by that region(compact clusters). The number of clusters,k, is always less than or equal to the number of observation, $k \leq n$.
In order to make distance between the kernel estimator and true density approximatlly very small with respect to h, the *Integrated Squared Error* is used to measure that as follows

$$ISE\{\hat{f}_h(x)\} = \int \{\hat{f}_h(x) - f(x)\}^2 dx \tag{1.2}$$

In this paper we are more interested in studying clustering based on kernel density estimation. An obvious way to do this is to partition the sample space using local minima defined by $\hat{f}'_h(x) = 0$. It will be interesting to examine the optimal bandwidth, h, which provides high coincidence with natural clusters, by measuring an agreement between partitions (clusters).

2 Agreement of partitions

By applying the method of kernel density estimation, different numbers of modes (clusters) can be yielded corresponding to various values of bandwidth. This variation may be useful to discover the characteristics of the data set structure. However, unsuitable selection of h leads to unnatural clustering results.

To avoid the risk of incorrect estimated structure of the observations, more attention should be given to the *validity* of a clustering algorithm at different values of smoothing parameter.

Therefore, the aim of this section is to measure the similarity between partitions, by using results that are obtained from clustering technique of the same objects, to determine number of paths (clusters) and assess bandwidth of such paths.

Now, let us suppose that there are n-objects in the set, $S = \{X_1, \cdots, X_n\}$,and any two partitions of S, say objects, $P_1 = \{P_{11}, \cdots, P_{1R}\}, P_2 = \{P_{21}, \cdots, P_{2C}\}$. Matching individuals, let n_{ij} be the number of objects placed simultaneously in P_{1i} and P_{2j}, can be obtained by cross-tabulating (matching matrix), $[n_{ij}]$, where

$$n_{ij} = \sum_{l=1}^{n} I[X_l \in P_{1i} \cap X_l \in P_{2j}], \quad (i = 1, \cdots, R, \quad j = 1, \cdots, C)$$

In fact, any two partitons are completely coincident if and only if every cluster in P_1 is a cluster in P_2, which only occurs if $R = C$.

2.1 The Rand index for similarity

Rand (1971) proposed a criteria which relies on measuring the coincidence of two different classifications with the same objects of the data set, this criteria is called the Rand index, I_R. Rand based his standard on: two objects are together in same clusters of P_1 and P_2; they are in different cluster of P_1 and P_2. Simply, it represents an agreement between P_1 and P_2 as follows

$$\{(X_l, X_{l'}) \in P_{1i} \cap (X_l, X_{l'}) \in P_{2j}, \quad 1 \leq l \neq l' \leq n\}$$

or

$$\{(X_l, X_{l'}) \notin P_{li} \cap (X_l, X_{l'}) \notin P_{2j}\}$$

From this, a measure of association for similarity of any two clustering methods of the same objects is defined (for $n \geq 2$) by

$$I_R = \left[\binom{n}{2} + T - \frac{1}{2}P - \frac{1}{2}Q \right] \Big/ \binom{n}{2} \tag{2.3}$$

where

$$T = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(n_{ij} - 1) = \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - n_{ij}$$

$$P = \sum_{i} n_{i\cdot}(n_{i\cdot} - 1) = \sum_{i} n_{i\cdot}^2 - n, \qquad Q = \sum_{j} n_{\cdot j}(n_{\cdot j} - 1) = \sum_{j} n_{\cdot j}^2 - n$$

$\binom{n}{2}$ represents the total number of both agreements and disagreements $(i.e\{(X_l, X_{l'}) \in P_{1i} \cap (X_l, X_{l'}) \notin P_{2j}\})$.

I_R could be obtained for any number of clusters or type of clustering. This index falls between zero and one: it is zero when the outcome of two partitions are completely different (no agreement); it is one when two partitions are completely identical (total agreement). Moreover, it has probabilition interpretation.

2.2 Rand index corrected for chance

Rather than measuring similarty between clusterings by using I_R, a corrected index is prefered when the expectation of the index does not take some constant value (e.g. zero) under an appropriate null model for contingency table. A general formula of the corrected index is

$$\frac{Index - Expected \quad index}{Maximum \quad Index - Expected \quad index} \tag{2.4}$$

The Rand index corrected for chance, $I_{R(adj)}$, was introduced by Hurbit and Arabie (1985). By assuming a maximum Rand index of 1.

3 Simulation design and results

In this study we generated groups (clusters) from $k = 3$ univariate normal distributions with different mean and same variance, so that each cluster possesses 100 observations. These three groups are combined to form a data set of size 300 and are refered to as *true clusters*. Besides this, we also redefine these groups by using points of local minima of the true density in terms of mixture normal ditribution of the above simulated data. To distinguish these new clusters from true clusters, they will be called *density clusters*. The reason for this is explained below. For each simulated data set, a kernel density estimate for a selection of smoothing parameters, $(h_1, \cdots, h_s)$, is evaluated to estimate a number of the modes (clusters), $m(h)$, and $ISE(h)$ for each h. Similarly, the indices will be used to compare the true clusters, used in model simulation, with clusters obtained from $\hat{f}_h(x)$, and they denoted by $I_R(h)^T$ and $I_{R(adj)}(h)^T$, and those that used in case of density cluster, $I_R(h)^D$ and $I_{R(adj)}(h)^D$ are copmuted as well. An average of all above terms is calculated, for simulated data set that are generated l times, where l=100, to obtain consistent result. All averages will be assessed in term of h^γ, h^β, h^α, where
(i). h^γ is such that, $\overline{m}(h^\gamma)$ is equal to the number of modes of the underlying density.
(ii). h^β is the value of h which minimizes average ISE, $\overline{ISE}(h^\beta)$
(iii). h^α are the smoothing parameters that maximize average indices, $\overline{I}^\bullet_\bullet(h^\alpha)$.

Table 1: Comparison of three smoothing parameters for various values of similarity

Index	$h^\gamma = 0.6101$	$h^\beta = 0.4830$	$h^\alpha = 0.6101$
$\overline{I}^T_R(h)$	0.8696	0.8677	0.8702
$\overline{I}^D_R(h)$	0.9402	0.9353	0.9402
$\overline{I}^T_{R(adj)}(h)$	0.7092	0.7000	0.7092
$\overline{I}^D_{R(adj)}(h)$	0.8679	0.8530	0.8679

From Table 1, we note that the values of the indices for h^γ and h^β are dissimilar in that h^γ yields higher values of similarity than h^β. Therefore, it would be better to say that kernel clustering is better for retrieving the structure of the object set when $h = h^\gamma$ than h^β. The values of $\overline{m}(h)$ and $\overline{ISE}(h)$ for these smooting parameters are: $\overline{m}(h^\gamma) = 3$, $\overline{ISE}(h^\gamma) = 0.0019$ about 11% larger than obtimized value, $\overline{m}(h^\beta) = 3.25$ and $\overline{ISE}(h^\beta) = 0.0017$, which seems to be slightly different from each other. In these simulations we find $h^\gamma = h^\alpha$ for all indices except $\overline{I}^T_R$ when h^α equal to 0.5467, the vast majority of the indices are completly the same as the indices for

h^{γ} execept when $h^{\alpha} = 0.5460$ which results in different values comparing with h^{γ} and h^{β}. Althogh, this value provides $\overline{m}(h^{\alpha}) = 3.07$ which not equal to the actual number of modes for the f ,it has higher indicator for coincidence than h^{γ} and h^{β}. In addtion, the density clusters supplied superior values of agreement with $\hat{f}_h(x)$ than true clusters whatever the smoothing parameter is.

In general, it seems from the results that are obtained, all smoothing parameters that we are interested in gives slightly different results. Because of this, the optimal value of smoothing parameters that researcher seeks depends on the aim. In our study we were look for the smoothing parameter that can provide the highest agreement between natural groups and kernel clusters. Moreover, we found that the best similarity can be achieved when the kernel clustering compare with density clusters. The reason for this is because both methods use the same way of partitioning the sample space (using local maxima and minima).

References

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, **2**, 193-218.

Rand, W. (1970). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, **66**, 846-850.

Wand, M.P. and Jones, M.C. (1990). *Kernel Smoothing*. Chapman and Hall, London.